R E P O R T R E S U M E S

ED 003 916 48

LEXICOGRAPHICAL PROBLEMS IN PASHTO.

BY- CHAVARRIS-AGUILAR, O.L. PENZL, HERBERT
INDIANA UNIV., BLOOMINGTON

REPORT NUMBER NDEA-VI-16 PUB DATE 60
EDRS PRICE MF-$0.09 HC-$0.72 18P.

DESCRIPTORS- DICTIONAIRES, *LEXICOGRAPHY, *PROBLEMS, *PASHTO, *DIALECTS, *DIALECT STUDIES, LANGUAGE RESEARCH, BLOOMINGTON, INDIANA

PROBLEMS IN DEVELOPING SUITABLE DICTIONARIES OF PASHTO WERE REPORTED. PASHTO WAS DESCRIBED AS A MEMBER OF THE IRANIAN BRANCH OF THE INDO-EUROPEAN LANGUAGES SPOKEN IN EASTERN AND SOUTHERN AFGHANISTAN AND IN NORTHEASTERN PAKISTAN. FIVE TOPICS DEALING WITH THE PASHTO LANGUAGE WERE DISCUSSED--(1) STATISTICS, (2) WRITTEN CORPUS AND LEXICAL STUDIES, (3) DIALECTS, (4) FOREIGN LOAN WORDS, AND (5) SUCH NONLEXICAL PROBLEMS AS TRANSLITERATION, TRANSCRIPTION, AND GRAMMATICAL IDENTIFICATION. THIS REPORT IS ONE OF A SERIES OF 13 PAPERS PRESENTED AT A CONFERENCE ON LEXICOGRAPHY, INDIANA UNIVERSITY, NOVEMBER 11-12, 1960. (GC)

ED003916

Title

LEXICOGRAPHICAL PROBLEMS IN PASHTO

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

I. Introduction

1. Statistics

Pashto is, along with Persian, Kurdish, Balochi, Ossetic, a member of the Iranian branch of the Indo-European languages and is spoken in Eastern and Southern Afghanistan and in Northwestern Pakistan, and in the tribal areas between these two countries. The number of speakers can be estimated at close to ten million, divided about equally between the two countries. Some of the published estimates as to numbers of speakers of Pashto, e.g. 4,000,000 in Meillet-Cohen and elsewhere, are much too low: the census of Pakistan of 1951 lists about 5,000,000 in that country alone, and in the border areas nominally under its control, who use Pashto as their mother tongue. On the other hand, some of the estimates coming from official and semi-official sources in Afghanistan may be considerably too high: 60% of an estimated total population of thirteen millions, i.e., 7,800,000, have been estimated to be speakers of Pashto just in Afghanistan.[1] In D. N. Wilber's Afghanistan (1956), in the Country Survey Series, the number of 'Pushtu' speakers is given as 4,500,000, and the total number of Pashto speakers in the world as 10,000,000; this would seem to us to be a reasonable figure. Incidentally, the great variety of spellings for the language, i.e. Pashto, Pakhtō, Pushto, Pushtu, Pukkhto, etc., is partly due to dialectal variation in the treatment of the second consonant (always written ښ),

---

1. Rahman Pashwak, Aryana. (1952) pp. 112 ff

and partly to difficulties and differing conventions in transliteration. (See GP, § 1.1 note.)[1] The first is the one preferred here.

2. Pashto in Afghanistan

That Pashto in Afghanistan has, in the past three or four decades, begun to experience a remarkable renaissance is due in some measure to the steps taken by former King Amanullah to raise Pashto to the status of the official language of the country and to the determined efforts of a relatively small number of dedicated Pashtun leaders and writers. But Persian with its greater literary prestige is, beside Pashto, still the official language of Afghanistan now. Persian is the language of the Royal Afghan Court. It is also the language of the capital of the kingdom, Kabul, where all the ministries and the only university--as well as the four or five most important secondary schools and special schools - are located. Before 1920 only Persian was used in Afghan schools, now Pashto is the medium of instruction in those primary and secondary schools situated in areas where it is the language of the majority of the population. In predominantly Persian-speaking areas it is taught like a foreign language from the third grade on up. Pashto is used in the press and on the radio--both government-controlled--and there is a Pashto Academy in Kabul, also government sponsored, which is concerned with publications, and research in Pashto. Traffic- and street-signs in Persian-speaking Kabul are in Pashto now. (See GP § 5.3)

3. Pashto in Pakistan

The position of Pashto in the new 'one-unit' province of West

---

1. GP = Herbert Penzl, A Grammar of Pashto: A Descriptive Study of the Dialect of Kandahar, Afghanistan (ACLS, Washington, D.C., 1955).

Pakistan is none too secure, although it is the native language of the President of the Republic Field Marshal Ayub Khan. Pashto is taught in the elementary schools of West Pakistan's Pashto-speaking areas and is an elective subject in its secondary schools, e.g. at Islamia College in Peshawar, where Urdu and English are compulsory subjects and where the medium of instruction, barring a few subjects taught in Urdu, is English. While Urdu has been pushed practically since independence as the sole medium of instruction in the elementary and secondary schools through-out West Pakistan, there has been increasing pressure for the use of Pashto in the former Northwest Frontier Province and there are signs that, as texts and teachers become available, Pashto will replace Urdu as the medium of instruction in the Pashto-speaking areas. Only this year the University of Peshawar instituted an M.A. program in Pashto. A city of the size and importance of Peshawar, the cultural center of the Pashto-speaking area and the former capital of the NWFP, has no daily newspaper entirely in Pashto. The number of publications by Pashto writers (prose fiction, poetry, magazines), however, has considerably increased since the beginning of the 'Pashtunistan' conflict between Pakistan and Afghanistan about ten years ago.

II. Written Corpus and Lexical Studies

1. The Written Corpus of Pashto

The earliest Pashto manuscript dates from the seventeenth century. Its text, and other texts with an allegedly even earlier date of composition do not give any evidence for morphological or phonological stages of Pashto earlier than those attested by the modern

language. Thorough lexicographical investigation of these texts has yet to be performed; only then will it become clear how their lexicon differs from that of the modern Pashto dialects. The lack of variety in the character of the literary monuments preserved to us and the relative meagreness of the textual tradition handicap any systematic historical investigation of the Pashto lexicon. There is, however, still a considerable corpus of unedited (and largely ignored) manuscript material in the India Office Library in London and a more modest collection at the Pashto Academy in Peshawar.

While a listing of the total annual literary production in Pashto from both sides of the Khyber Pass still would not offer a seriously taxing problem to bibliographers, the modern written corpus of Pashto is actually of considerable magnitude, for newspapers and magazines have appeared for decades in the cities of Afghanistan.

2. Lexicographical Studies in Pashto

The lexicographical work done in Pashto consists of:

1) two dictionaries, now out of print: Major H.G. Raverty, Dictionary of the Puk'hto, Pus'hto, or Language of the Afghāns (London, 1860; 2nd ed, 1867); Henry Walter Bellew, A Dictionary of the Pukkhto or Pukshto Language (London, 1867; 2nd ed., Lahore 1891). Both of these dictionaries, as well as grammars, textbooks, and text editions, resulted from the efforts of the British administration to encourage proficiency in the native languages of British India. Assistant Surgeon Bellew added also a 'reversed

part" (Preface, p. XI), English-Pashto, but the Pashto-English part, 'aided' by Raverty's work, is his main contribution, mostly derived from his reading of 'Pukkhtūn authors.'

2) Major George Waters Gilbertson, The Pakkhto Idiom: A Dictionary (Hertford, 1932) in two volumes, contains a collection of Pashto sentences and their translations alphabetically arranged under one English key word or idiomatic phrase that has an equivalent in the Pashto sentences. They must have been collected from the three assistants (Arif Ullah, Yūsufzai; Makhmūd, Afrīdī; Alī Akbar Khān, Qandahārī) mentioned on the title page. A comprehensive New English Pakkhto Conversational Dictionary, announced by the author in the Foreword (p. IX) apparently never appeared in print.

3) A Pashto-Russian and a Russian-Pashto dictionary: P.B. Zudin, Kratkiĭ Afgansko-Ruskiĭ Slovar' (Moscow, 1950; 568 pp.) containing 12,000 words and appended brief notes on Pashto by E.A. Bertel's; P.B. Zudin, Russko-Afganskiĭ Slovar' (Moscow, 1955; 1176 pp.) containing about 21,000 words and a brief grammatical sketch by D.A. Shafeev (pp. 1035 ff).

4) There have been a variety of dictionary compilations published in Kabul, most of which are no longer available, even in the libraries of Kabul, e.g. Paštō Sind (1316 A.H.), by Muhammad Gul Muhmand, 266 pp. Pashto-Persian; Paštō Qāmūs (1330-3 A.H.), by Sadiqullah Rištīn; Afghān Qāmūs: Farsī ba

Paẋtō, by Abdullah Afghānī Nawīs, 3 vols, Pashto Academy, Kabul 1957. In Nawīs' (Nawees') dictionary Persian entries are followed by Persian glosses and Pashto equivalents. The Pashto Academy at the University of Peshawar is engaged in a monumental Pashto-Urdu-English historical and dialectal lexicon project. At this moment they are halfway through the 'Alif,' the first character of the Perso-Arabic script which Pashto employs.

The etymologies of Pashto words were treated in George Morgenstierne's An Etymological Vocabulary of Pashto (Oslo, 1927).

III. Pashto Dialects in Field and Texts

1. Dialectal Variations in Pashto

Pashto shows a great variety of dialectal subtypes. Sir George Abraham Grierson's Pashto specimens in vol. 10 of the Linguistic Survey of India provide only a very inadequate sampling of the dialectal distinctions. They are phonemic and phonetic, morphological, and lexical. All dialects of Pashto seem to show the same syntactic structure and certainly the same grammatical categories. Needless to say, only extensive and systematic field work will reveal in detail the exact distribution of the dialects of Pashto, a task to date quite neglected. For our purposes it will be sufficient here to illustrate the three types of dialectal distinctions. The phonemic differences are the most striking. They involve differences in the phonemic pattern; the presence or absence of certain allophones; the distribution of the phonemes and their frequencies. In the dialect of Kandahar--which agrees more than any other dialect with the generally accepted

orthography for Pashto—retroflexed sibilants, both voiced and voiceless, appear medially in, e.g., kēṣī 'becomes' and in paṣtō 'Pashto', respectively; there occur likewise a voiced alveolar affricate and a voiceless alveolar affricate, e.g. in dzam 'I go' and in tsōk 'who', respectively. The dialect of Warḍak, south of Kabul, has the same number of phonemic contrasts, but instead of the retroflexed sibilants a contrasting pair of voiced and voiceless medio-palatal fricatives occur, e.g. ǰēǰī and paçtō. Most dialects of the province of Kabul retain the medio-palatal voiceless fricative in paçtō, but the medial consonent in kēṣī has coalesced with the voiced velar stop /g/; kēgī. In the dialect of the Peshawar area as well as in some of the Eastern dialects not one of the four phonemes mentioned above for the Kandahar dialect occurs; they have coalesced, the voiced retroflexed sibilant with the voiced velar stop, as in kēzī; the voiceless retroflexed sibilant with the voiceless velar fricative, as in paxtō; the affricates have been simplified and have coalesced with /z/ and /s/, as in zam and sōk. Worthy of note among the characteristic allophones of some of the dialects are the high allophones of the two mid vowels /ē/ and /ō/ before a following vowel of high tongue position, to the point where it is often difficult to distinguish clearly between /ē/ and /i/ and between /ō/ and /u/: e.g. /ē,/ in kēwī, /ō/ in gōrū 'we see'.

Distributional differences of phonemes concern particularly the sibilants, where the dialects show varying distribution of /s/ and /š/ and also of /z/, /ž/ and /ǰ/ in some words of high frquency. In many dialects of the provinces of Kabul and Paktia, in Ghazni, Kandahar, and Herat in Afghanistan and in Dera Ismail Khan in Pakistan, šta 'there is'

and ɣaɣ 'I can' are regularly heard instead of x̌ta and x̌aɣ. For x̌aba 'language, tongue,' some central dialects have ɣaba and some Eastern dialects in the Kunar Valley have x̌eba.

Morphological distinctions between the dialects concern, e.g., some substantive and verbal endings: kōrō obl. pl. in Kandahar for kōrúnō 'houses'; -ast for -ay as the second person plural ending of the past tense and some present forms in Kandahar, Farah, Herat and Dera Ismail Khan, e.g. yāst 'you are'. The third person singular masculine of the past of the verb can be -O ('zero'), -a, -ā, -aX, e.g. walid, walide, walidō (Peshawar), walidey (Kandahar) 'seen.' It is interesting to note that, with few minor exceptions, Afghan grammarians in their textbooks describe as models for instruction the forms of their own dialects.[1]

Lexical distinctions reveal dialect differences of considerable extent, e.g. such items as dōbay and ōṛay 'summer,' lōy and stər 'great, big,' tsōṛmunay and psarlay (also sparlay) 'spring,' lmar and starya (de lmar starya) 'sun.'

2. The Dialects and the Dictionary

One problem for the lexicographer is to what extent dialectal variation should be included in his lexicon and how it is to be treated. For Pashto this is no great problem. There is a more or less standard

---

1. Herbert Penzl, Afghan Descriptions of the Afghan (Pashto) Verb, JAOS, 71 (1951), 97-111; Die Substantive des Paschto nach afghanischen Grammatiken, ZDMG, 102 (1952), 52-61.

orthography accepted throughout the Pashto-speaking world, an orthography which represents, for the consonants at any rate, the maximum phonemic distinctions as found in the dialect of Kandahar. Thus phonemic variation can to a large extent be ignored. The one complicating factor--and it is not too serious as yet--is that there seems to be arising a tendency among some writers to model the orthography in some respects after their spoken dialect.[1] Only a glossary based on one or a few specific texts can, however, include such variations as, e.g., ښځه for ښڅه 'woman,' and hyperforms such as دمې for مې 'my,' even though their general textual frequency may be fairly high. Dialect dictionaries and glossaries, of course, would have to render the specific phonemic forms in question as do, e.g., Lorimer's Waziri study,[2] and the Afridi word lists in The Modern Pushtu Instructor (Peshawar 1943, vol. II) by Qazi Rahimullah Khan.

The morphological variants such as those mentioned above present no problems, since they are not the concern of the dictionary but rather that of the descriptive and reference grammars of the language. The lexical variants which represent dialectal differences whill have to be included if they occur in texts, their inclusion to be determined primarily by their textual rather than by their field frequency. This is true of such items as ستر besides لوی, two morphemically entirely different lexical items, as well as of morphemic variants (which may really in some cases represent phonemically different dialectal types),

---

1. D.N. MacKenzie, A Standard Pashto, BSOAS, 22 (1959), 231 ff.
2. J.G. Lorimer, Grammar and Vocabulary of Waziri Pashto (Calcutta 1902).

such as lyār beside lār 'road,' lund beside lund 'wet,' wrəḍz beside wrəḍz 'day,' mṛaka beside mṛaka 'earth'. In morphology and lexicon there is very little difference between written and spoken Pashto within a single dialect or group of dialects. There is certainly nothing comparable to the 'diglossia,' to use Ferguson's term, that is prevalent in Afghan Persian, where a tremendous gulf separates the literary Persian of Afghan publication from the spoken 'Kabuli' of daily discourse.

The Pashto Academy in Kabul has attempted in writers' conferences sponsored by it to establish morphological and sometimes morpho-lexical standards for writers on both sides of the Pakistan-Afghan border.[1] It has upheld the standard orthography. Phonetic spellings such as ʒəba ښ or ɣəba ښ for ʒəba; the frequent Peshawar ې (ē, actually /ɛ̄/ contrasting with /ə̄/) for /əy/ in kawalay were rejected. The zero ending for the 3rd asc. sing. preterite form, e.g. wəlid 'seen', which is quite rare in the dialects, was recommended instead of -ə, -ā, -əy; the Kandahar -āst beside -əy for the 2nd pers. pl. of the verb was approved.

3. Textual and Field Frequencies

We mentioned earlier that there is no great difference between spoken and written Pashto. A writer will, therefore, while making concessions as to phonemic shape by adopting the 'standard' shape as against that of his own dialect, otherwise feel free to make use of most of the lexicon of his own dialect in writing. This seems to imply for

---

1. See M.H. Kāker, in Kābul, issue 465 (1958), pp. 1-14.

the lexicographer that the Pashto lexicon can be gotten either by means of a quasi-philological analysis of texts or through field-work with informants. The character of the available written corpus of Pashto is redundant to a considerable extent, thus limited in its total range. Some items will have a very low (or even zero) textual frequency, where the field frequency can be shown to be quite high.

The conclusion to be drawn would seem to be that in the case of Pashto a combination of textual analysis and field-work will yield the lexicon essential for a dictionary that is to contain all items of high textual and field frequencies.

IV. Foreign Loanwords in Pashto

A special lexicographical problem arises from the presence in Pashto of many Persian, Arabic, Indic (Urdu) and western terms. Bellew already stated in his Preface (p. VII), 'I have experienced considerable difficulty in deciding upon the words of foreign origin--principally Arabic and Persian, but still of common use in the Pukkhto-- that should find a place in these pages'.

The number of foreign words will differ in texts from different areas. Words with high textual and field frequencies have to be included. In Afghanistan a number of words that are puristic or artificial creations sometimes intended to replace established Arabic-Persian terms may have a higher text-than field frequency; e.g. ژوندون versus maktab 'school'. No term with a relatively high textual frequency can be excluded.

Some western terms show a morphemic variation that is not always represented by the Perso-Arabic orthography; e.g., gēr and gīr 'gear,' klīnar and kilīnar 'driver's assistant.' Lexical listing in such cases will have to include both possibilities. The reason for this morphemic variation is the increasing adaptation to native speech patterns by some, and the puristic attempts at remodelling by other speakers familiar with the languages from which the borrowings came. Western loans are found in the field of politics, science, western culture, technology, motor transport, etc. Speakers familiar with especially English, French, and German brought these words into the language. Sometimes the contact with specific languages will determine the particular shape of the loans; e.g., sāying is competing with sayāns, English and French 'science,' Āstriā with Ōtriš 'Austria.' The role that English is playing in secondary education favors the spread of loanwords with an English base. The textual frequencies of some of these words is low but their field frequencies high: e.g., tšēn '(snow) chain,' sīt 'seat' (of bus or car), bēlt 'belt.' etc. The dictionary will have to include all Arabic, Persian, western and other terms whose frequencies justify it..

V. Neulexical Problems in Pashto

1. Transliteration and/or Transcription

The items collected from texts will, of course, be given in the standard Arabic orthography. As stated above, deviant phonetic spellings employed by writers of a dialect with a phonemic system different from that of the 'standard' language as represented by the generally accepted orthography, can only be included in special glossaries for specific texts or dialects. The same is probably true for occasional spellings

that offer colloquially general or at least very wide-spread replacements of, e.g., p for f, k for q, etc., in Arabic and Persian loans. If, however, their textual frequency in some lexical items should prove to be sufficiently high, they may have to be included as morphemic variants (cf the type mraδz and mrəδz as mentioned III. 2).

On the Pashto-English side of the dictionary every entry in the Arabic alphabet must be followed by an indication of its standard phonemic shape by means of a Roman alphabet notation. This is very necessary because the Arabic orthography does not indicate the short vowels a ə i u nor does it make any distinction between o ū u ai əi ui əu, which are all written as 'wāw' (و).[1] The practices of the extant Pashto dictionaries (see II. 2 above) show some difference in the Romanized writing of Arabic (and Persian) loanwords. The entries in the dictionaries by Raverty, Bellew, Gilbertson transliterate the Pashto letters not only by such symbols as f (ف), q (ق), ʿ (ع), h (with a diacritic) for ح , but distinguish ت and ط ,ث ص س ,ز ,ذ ظ ض by t̤ s̤ z̤ symbols with differentiating diacritics. This practice, incidentally, was continued in GP (see § 2.5), where diacritics are only used for optional transliteration, when their omission would not change the phonemic representation. Bellew gave in the English-Pashto part of his dictionary only the Roman transliteration of the Pashto entries, which are however 'reversible' into the Arabic orthography. In Zudin's dictionaries the Cyrillic transcription includes only f, but no special symbols for ك (written k), the glottal stop, and only one symbol each

---

1. L. Herbert Penzl, Orthography and Phonemes in Pashto (Afghan), JAOS. 74 (1954), 74-81.

for /t/ /s/ /z/.

In spoken Pashto in a varying degree (see GP §§ 37-40) the phonemes /ṭ/ /q/ /ʾ/ sometimes /h/, e.g. h with pharyngeal friction, are used by literate speakers in a more formal style of speech. Should the dictionary indicate the formal pronunciation, which would agree with the Arabic spelling, should it substitute a colloquial transcription or should it indicate both formal and colloquial pronunciations? Since the colloquial pronunciation can be derived from the formal and elegant one by some simple rules of transformation, but not vice versa, it will seem preferable to indicate consistently the formal pronunciation. However, with some lexical items of high field-frequency a colloquial pronunciation may have to be added to the formal one.

Although there is no clearly defined standard of pronunciation for Pashto, the selection of a 'standard' for the transcribed (or transliterated) entries will have to be determined by the phonemic distribution in the main dialectal types of Pashto, which, on the whole, agree quite well, at least in the case of the vowel phonemes. Zudin selected for his dictionary a Kandahar pronunciation (see II.2).

2. The Grammatical Identification of Entries

A dictionary should include the grammatical information necessary to identify the word class and the essential morphological features of an entry. Text and field occurrences involve not just the 'quotation' form. The Pashto word class will customarily be identified by the corresponding word class of the English gloss or definition. The morphemic subclasses can be indicated by either some additional forms, from

which the remaining inflectional features can be derived, by suitable code symbols, if brevity is called for, or by a combination of both. In the glossary to the GP, for example, a combination of both practices was used to characterise the lexical items. For substantives the gender and plural-formation, for adjectives the formation of their feminine, for verbs the formation of their two past tenses (I and II) were considered the principal criteria for their subdivision into morphemic subclasses. Thus substantives would be marked m(asculine) or f(eminine) and their plural and oblique forms indicated by a numeral: m 1, m 2, m 3, m 4, m 5; f 1, f 2, f 3, f 4, f 5, f 6. For instance kĭr 'house' would be labeled m 1, which implies a plural kŏrúna (also obl. sing. II kŏra, obl. pl. kŏrŭ); ẍpa 'night' is f 2, which indicates the plural ẍpe (also obl. sing. ẍpa, obl. pl. ẍpe). The adjective uẍd is marked adj. II, which implies a feminine uẍda, plus a whole set of other predictable endings. Five adjective classes are differentiated altogether.

The verbs were labeled Verb I, II, III, IV, IV-A: e.g., tarəm 'I bind' verb I, which means a predictable formation of the past tenses with no change from the present stem; rasēzəm (past rasēd-) 'I arrive, reach', verb II, with the change in the past stem as indicated; the same way with gŏram (past kat-), 'I look at,' verb III, where present and past stems seem suppletive. Only seven verbs were marked verb IV and the stems of present II, past I and past II had to be indicated; e.g., kawəm (pres. II [wə] kəm, past I kaw-, past II [wə] k-) 'make, do' verb IV. Verbs of class IV-A are really compounds of kĕẍem and kawəm (2 class IV verbs) and 4 forms have also to be indicated there.

The principle of indicating, if space permits, all morphological features of entries not predictable from the shape of the quotation form (with a minimum of grammatical knowledge) seems sound.

O.L. Chavarria-Aguilar
Herbert Penzl

University of Michigan
Research Project in
Near and Middle Eastern Languages